How many grains are needed for quantifying catchment erosion from tracer thermochronology?

Andrea Madella[1], Christoph Glotzbach[1], Todd A. Ehlers[1]

[1]Department of Geosciences, University of Tuebingen, Schnarrenbergstr. 94-96, 72076 Tuebingen, Germany

5 *Correspondence to*: Andrea Madella (andrea.madella@uni-tuebingen.de)

Abstract. Detrital tracer thermochronology exploits the relationship between bedrock thermochronometric age-elevation profiles and a distribution of detrital grain-ages collected from river, glacial, or other sediment to study spatial changes in the distribution of catchment erosion. If ages increase linearly with elevation, spatially uniform erosion is expected to yield a detrital age distribution that mirrors the catchment's hypsometric curve. Alternatively, a mismatch between detrital and

10 hypsometric distributions may indicate non-uniform erosion within a catchment. For studies seeking to identify the pattern of erosion, measured grain-age populations rarely exceed 100 grains due largely to the time and costs related to individual measurements. With sample sizes of this order, discerning between two detrital age distributions produced by different catchment erosion scenarios can be difficult at a high statistical confidence level. However, there is no established method to quantify the sample-size-dependent uncertainty inherent to detrital tracer thermochronology, and practitioners are often left

15 wondering 'how many grains is enough?'. Here, we investigate how sample size affects the uncertainty of detrital age distributions and how such uncertainty affects the ability to uniquely infer the erosional pattern of the upstream area. We do this using the Kolmogorov-Smirnov statistic as metric of dissimilarity among distributions, based on which the statistical confidence of detecting an erosional pattern is determined through Monte Carlo sampling. The techniques are implemented in a new tool (*ESD_thermotrace*) to consistently report confidence levels as a function of sample size and application-specific

20 variables. The proposed tool is made available as a new open-source Python-based script along with test data. Testing between different hypothesized erosion scenarios with this tool provides thermochronologists with the minimum sample size (i.e. number of bedrock and detrital grain-ages) required to answer their specific scientific question, at their desired level of statistical confidence. Furthermore, in cases of unavoidably small sample size (e.g., due to poor grain quality or low sample volume), we provide a means to calculate the confidence level of interpretations made from the data.

1 Introduction

Tracer thermochronology uses the distribution of single grain thermochronometric ages of detritus to infer the spatial pattern of erosion in the source area (e.g. Stock et al., 2006; Vermeesch, 2007). This approach is typically applied where bedrock age data exhibit a clear age-elevation relationship, allowing inference of the relative contribution of source elevations from the detrital grain-age distribution. A detrital grain-age distribution that mirrors the catchment's hypsometric curve (i.e. the

30   cumulative distribution function of elevation area), is generally interpreted as indicative for spatially uniform erosion. Conversely, a detrital age distribution skewed towards younger (or older) ages may be the consequence of focused erosion at lower (or higher) elevations (Spiegel et al., 2004). Tracer thermochronology has been shown to be a powerful tool to investigate the sub-catchment-scale variability of denudation. Geomorphologists have been able to infer changes in climatic parameters (Nibourel et al., 2015; Riebe et al., 2015), glacial erosional processes (Clinger et al., 2020; Ehlers et al., 2015; Enkelmann and

35   Ehlers, 2015), sediment dynamics (Lang et al., 2018), relief evolution (McPhillips and Brandon, 2010), occurrence of mass-wasting (Vermeesch, 2007; Whipp and Ehlers, 2019) and differences in rock uplift (Glotzbach et al., 2013, 2018; McPhillips and Brandon, 2010). Unfortunately, the number of measured detrital ages for tracer thermochronology is often dictated by inherent limitations of the sampled material and/or by available finances, rather than a science-based choice. Detrital sample sizes often range between 40-120 grains (Clinger et al., 2020; Ehlers et al., 2015; Glotzbach et al., 2018; Lang et al., 2018;

40   McPhillips and Brandon, 2010; Riebe et al., 2015; e.g. Stock et al., 2006; Vermeesch, 2007), and are considered to yield high-confidence results when surpassing ~100 grains based on previous work on sediment provenance analysis (Vermeesch, 2004). However, in unfortunate cases, discerning between two measured distributions generated from different erosional patterns can be statistically impossible at a high confidence level even with more than 100 grains. Although this issue is well-known to the community (e.g. Avdeev et al., 2011) since the early days of such detrital studies (Brewer et al., 2003), there is no established

45   method to quantify the sample-size-dependent uncertainty inherent to detrital tracer thermochronology, and the number of measured grains may often be chosen based on convenience and/or habit.

Here, we complement previous work by investigating how sample size affects the uncertainty of detrital cooling age distributions and the related confidence in addressing the erosional pattern of the upstream area. We discuss the approaches

50   used in previous case studies, based on which we develop a tool (Earth System Dynamics - *ESD_thermotrace*) to consistently report confidence levels as a function of sample size and case-specific variables. We illustrate our approach using a published dataset from the Sierra Nevada Mountains, California (Stock et al., 2006). The proposed tool is made available as a new open-source Python-based script along with test data. We demonstrate how *ESD_thermotrace* can assist future tracer thermochronology studies in defining the necessary sample size to answer their specific scientific question. In cases where

55   larger sample sizes are impossible to achieve, the study's confidence level can be quantified using our script.

2 Background information

Single grain detrital age distributions are extensively applied in classical detrital geochronology (Hurford and Carter, 1991), where U/Pb crystallization ages of zircon constitute by far the most used (and possibly overused) tool (Andersen, 2005; Malusà et al., 2013; Spiegel et al., 2004). In this type of application, the aim is to obtain the spectrum of all age components (i.e. age

60   peaks) that characterize a siliciclastic sediment. If a range of assumptions hold (Malusà et al., 2013), the provenance of a sediment sample's source area can be inferred by matching the detrital age components to those of known upstream geological units and/or events. For that purpose, the number of measured detrital grains determines the confidence of detecting

minoritarian age components. An exhaustive probabilistic method to report such confidence exists (Vermeesch, 2004) and is not the focus of this study. The absolute age components of the source area are in fact unimportant in detrital tracer thermochronology (Avdeev et al., 2011), for which monolithological catchments are best suited to avoid mineral fertility issues in the source rock (Fig.1). The focus here is the dissimilarity between the distribution of ages found in the source area and in the fluvial/glacial/hillslope sediment is derived from, regardless of their absolute age component. For this purpose, the uncertainty caused by a small sample size strongly limits the minimum dissimilarity that can be resolved between two distributions. This minimum resolution directly affects the power of our inferences about the upstream erosional pattern – commonly the ultimate scope of tracer thermochronology.
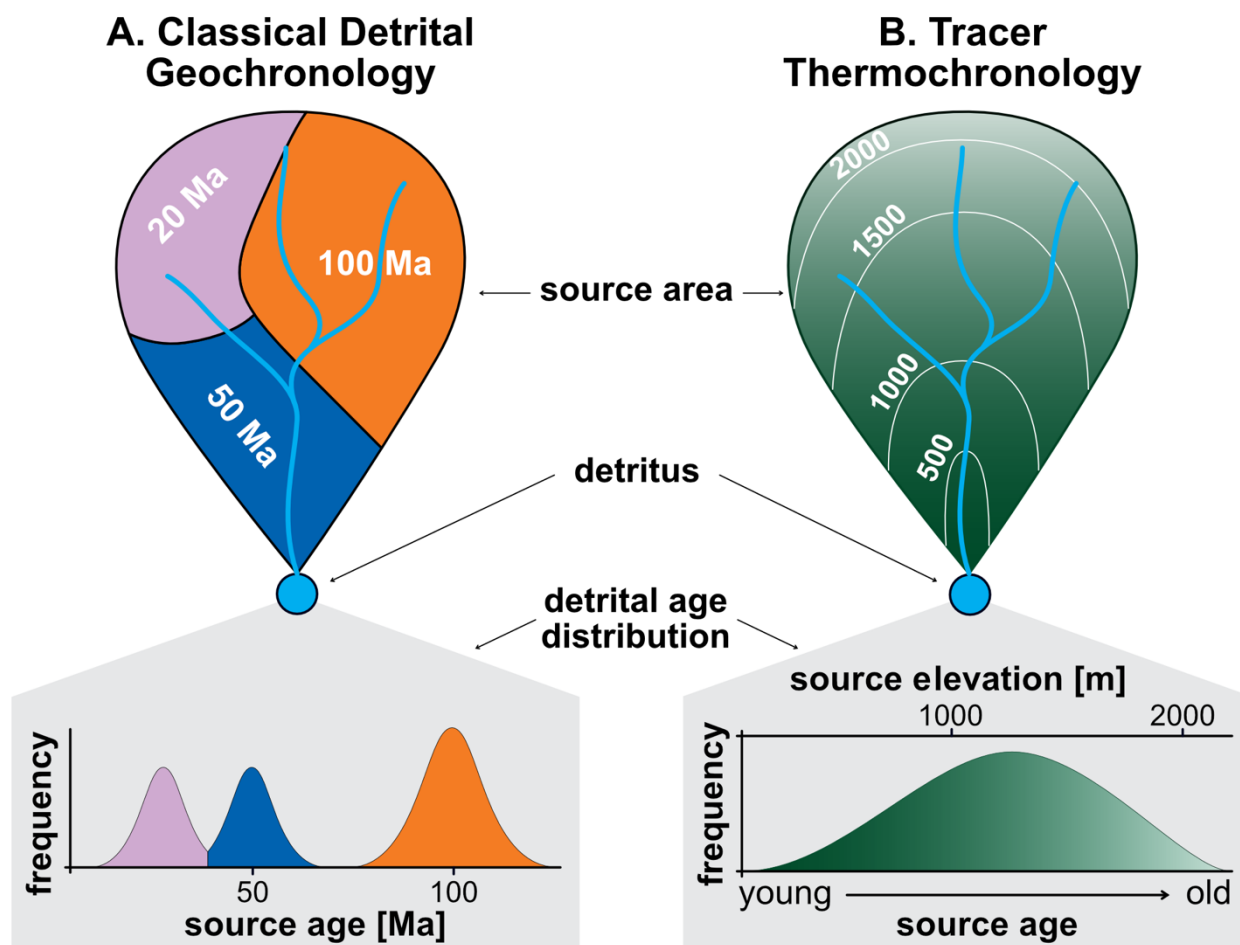

Figure 1 - Sketch of the difference between classical detrital geochronology (A) and tracer thermochronology (B). A: discrete age components are found in the detritus and refer to different upstream geological units. B: a continuous detrital age distribution informs the relative abundance of material sourced from different elevations, based on a known age-elevation relationship.

In the following we summarize the conceptual model concerning this matter and the approaches that have been used to address it thus far. Let us consider a monolithological catchment and a set of detrital grain-ages

75   measured at its outlet. The observed grain-age distribution should match a predicted distribution that is forward-modelled stacking the following layers of geologic information about the upstream area (Fig.2):

I.   The catchment hypsometry, or the distribution of the catchment's cumulative area as a function of elevation. The
80   hypsometric curve is constructed based on the digital elevation model (DEM) of the study area, and it has a negligible uncertainty.

II.   The bedrock age-elevation data, possibly a linear function based on a dataset of ages with associated uncertainty,
85   measured at known elevations.

III.   The mineral fertility, which informs the original abundance of the target mineral in the different elevation ranges and is mostly a function of lithologic variability. This is a critical parameter that can lead to gross misinterpretations if
90   ignored (Malusà et al., 2016).

IV.   A pattern of erosion across the catchment, which is the dependent variable of the system. This is any map of erosion probability, a scenario to be tested against uniform erosion.

95   The ultimate aim of tracer thermochronology is to find the best-fit pattern of erosion by minimizing the mismatch between observed and predicted distribution.

3 Comparing predicted and observed distributions

Predicted distributions should be constructed accounting for all above
100   information and related uncertainties, such that the confidence of the fit to the observed distribution also accounts for them. Brewer et al. , and successively Ruhl and Hodges (2005), were the first to compare the distribution of detrital thermochronometers to that of age-elevation data. Although the scope of their work differs from more recent tracer
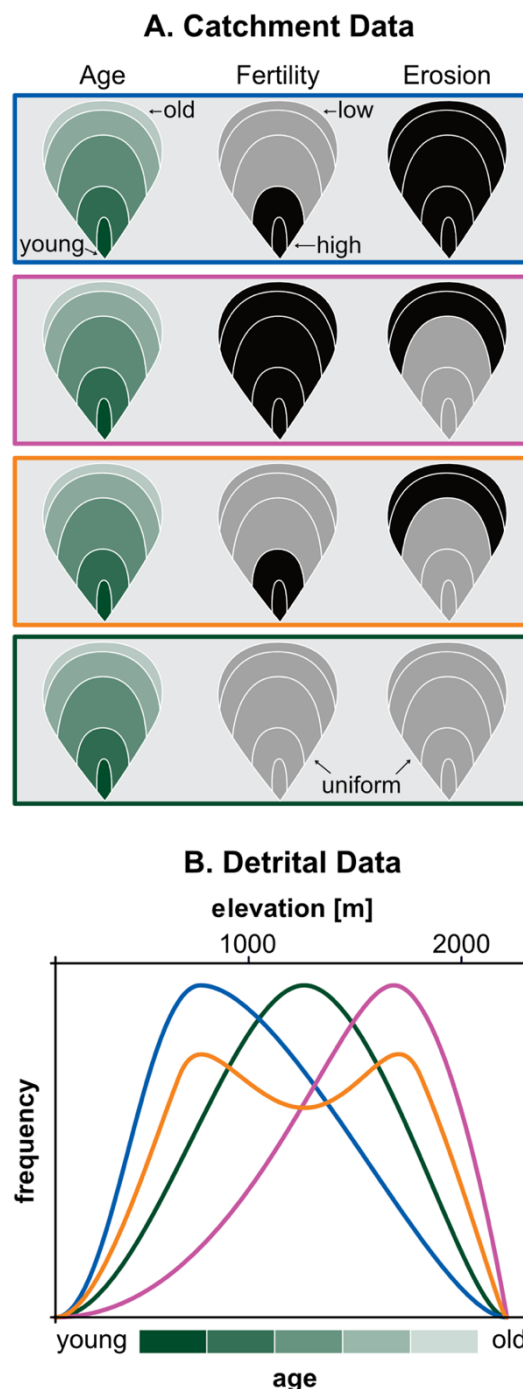
Figure 2 - A: The catchment of Fig.1B with known bedrock age (shades of green) is subject to 3 scenarios of spatially varying fertility and erosion. The box outlines refer to the curves below. B: detrital distributions obtained from the different scenarios in above. The green curve refers to spatially uniform erosion and fertility.

105 thermochronology, the evaluation of the dissimilarity between predicted/observed distributions remains the main object of their studies. These authors construct a synoptic probability density function (SPDF) of the observed data by "stacking" the gaussian distributions of all measured grain-ages, each with their analytical error (this is equivalent to the $SPDF_t$ in Table 1 of Vermeesch (2007)). In addition to the observed one, a predicted SPDF is also constructed with the same method, where the predicted grain-ages are a random subsample of the hypsometric curve and are each given an arbitrary average uncertainty.

110 Brewer et al. (2003) define the mismatch P_{diff} between observed and predicted SPDF with equation (1):

$$P_{diff} = \frac{\sum_{t=0}^{t=\infty} |P_1(t) - P_2(t)|}{2} \times 100, \tag{1}$$

where (P_1 and P_2) are the probabilities of the two distributions calculated at each age step (t). P_{diff} relates to the area comprised between two SPDF in the age-frequency space. Brewer et al. (2003) calculate the 95% confidence mismatch between observed and predicted SPDF through a Monte Carlo simulation. The same approach was complemented by Stock et al. (2006), who

115 performed a Kuiper equality test (Kuiper, 1960) for each Monte Carlo iteration and used the percentage of successful tests as confidence level.

Vermeesch (2007) has shown that, for the purpose of comparison, observed and predicted distributions are best expressed as a cumulative age distribution (CAD) rather than SPDF. A CAD is a step-function with the sorted mean ages on the x axis and

120 the related quantiles on the y axis (Vermeesch, 2007). This method is preferred because it avoids the possible sources of bias introduced by (i) choice of smoothing parameter in kernel density estimations (KDE), (ii) binning in histograms and (iii) uncertainty-based weighting in SPDF curves (Vermeesch, 2012). To evaluate the goodness of fit between observed and predicted CADs, Vermeesch (2007) uses the Kolmogorov-Smirnov (KS) statistic, which informs the maximum distance d_{KS} between two cumulative distribution functions as follows (Massey, 1951 and references therein):

125 $$d_{KS} = maximum \left| CAD_{observed}(t) - CAD_{predicted}(t) \right|. \tag{2}$$

Given an observed CAD with k observations (i.e. ages), d_{KS} is calculated for several $n=k$ sub-samples of the predicted CAD. The 95th percentile (d_{KS_95}) of all sorted d_{KS} is used to define a region of significance around the continuous predicted CAD. An observed CAD that plots entirely within the range $CAD_{predicted} \pm d_{KS\_95}$ (Fig.3) is considered to be drawn from the same predicted age population, at the 95% confidence level. As an alternative to this iterative method, the confidence region for a

130 predicted CAD can be approximated calculating the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality bounds as follows:

$$d_{KS\_95} \approx \varepsilon = \sqrt{\frac{\ln\frac{2}{\alpha}}{2k}}, \tag{3}$$

where the DKW distance ε approximates well d_{KS_95} as a function of the confidence level $(1 - \alpha)$ and the number of observed detrital grain-ages k (Fig.3) (Massart, 1990).

Riebe et al. (2015) further developed the bootstrapping approach described above to age distributions (in SPDF form). Instead

135 of basing their analysis on the KS statistic, the 95% confidence envelope of the prediction is iteratively estimated at each age step t_i. For each t_i, the distribution of 10,000 predicted age frequencies $SPDF_{predicted}(t_i)$ is used to draw 2.5th and 97.5th quantiles
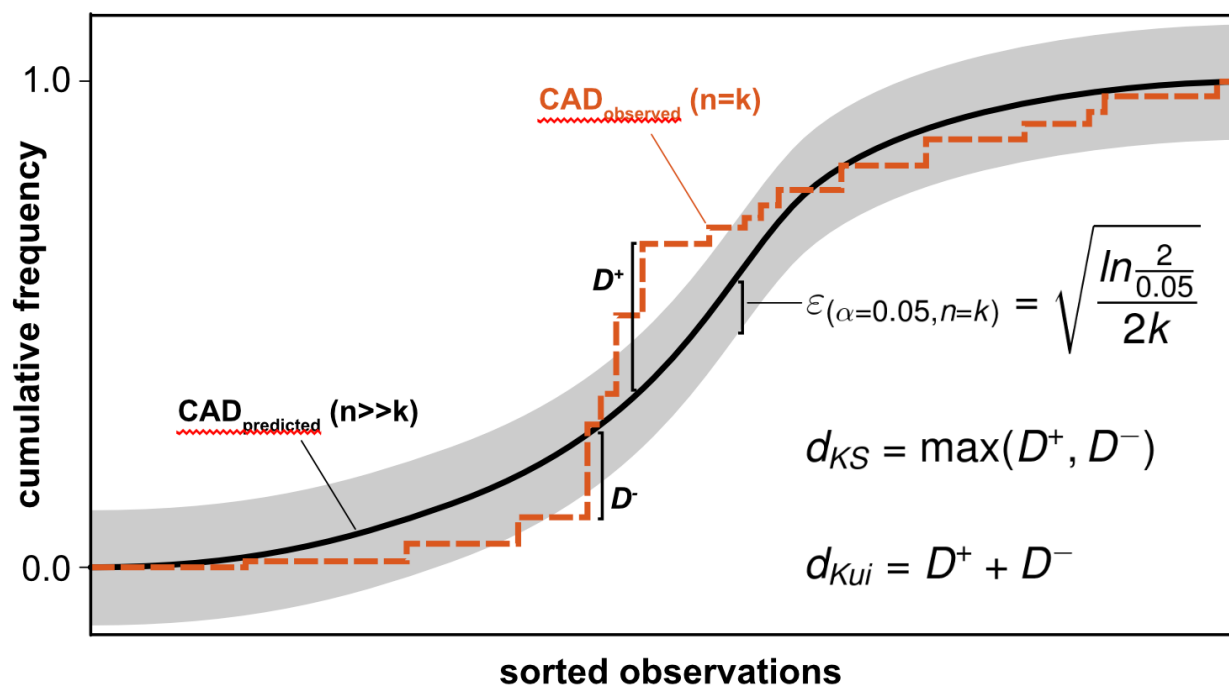
Figure 3 - Example of metrics to compare two cumulative distributions drawn from different sample sizes. Here, an observed Cumulative Age Distribution (CAD) for n=k (orange stepwise dashed line) is compared to a predicted CAD drawn for n>>k (solid black line). The Kolmogorov-Smirnov statistic d_KS equals the maximum absolute vertical distance between distributions in the observations-frequency space. The Kuiper statistic d_Kui equals the sum of both positive and negative maxima. For α=0.05 (i.e. 95% confidence) and n=k, a Dvoretzky-Kiefer-Wolfowitz confidence region can be calculated (gray shading). In this example, observed and predicted distributions are drawn from two statistically different populations at the 95% confidence level, because the orange curve exceeds the gray region.

at each age step. Because the age steps t_i relate to elevation steps in the catchment through the bedrock thermochronometric data, those elevation steps exhibiting excess (>97.5[th]) or deficit (<2.5[th]) frequency are interpreted to produce sediment in excess or deficit with respect to the reference scenario of uniform erosion, at the 95% confidence level.

140

The approaches summarized above are well-suited to determine whether an observed detrital age distribution is significantly dissimilar from a predicted distribution. Such predicted distributions have so far been constructed based on the source bedrock age-elevation data and assuming spatially uniform erosion. Because scenarios other than uniform erosion have not been used to make predictions, the inferred pattern of denudation in the cited studies is qualitative to semi-quantitative. Moreover, the

145 confidence level of the results in terms of dissimilarity is commonly assessed only after samples have been processed and dated. This approach, though very convenient and rather widespread in geologic studies, may undermine the statistical rigor of a study. A preferable approach is to define the confidence level for rejecting or confirming the study's hypothesis in the first place. The sampling strategy and the number of measurements and resources allocated to test the hypothesis should be decided therefrom. In the next paragraphs we introduce the new *ESD_thermotrace* software, a tool built on the cited previous

150 work, to help determine the number of grain-ages necessary to produce statistically sound results as a function of bedrock data, catchment hypsometry, erosion scenario and the chosen confidence level. We will also show the use of our tool to evaluate the fit between predicted and observed distributions.

4 ESD_thermotrace: A Python-based Jupyter Notebook to study the uncertainty of tracer thermochronology datasets

We developed a new tool to predict detrital age distributions, calculate their dissimilarity to the observations and determine
155 the confidence level of the results. The software performs the steps briefly outlined here below. For additional details, the reader is referred to the illustrative case study down below and the well-commented *ESD_thermotrace* code itself (Madella et al., 2021)[Please note that a temporal preview access to the cited future DOI landing page is available: https://dataservices.gfz-potsdam.de/panmetaworks/review/c375e8eb059bb94ace622f2075b99baf6c513d3b560989a6b307bbe5d5a20217/]:

1. Bedrock age map interpolation
160   o Input: bedrock age-elevation dataset and digital elevation mode.
         o Output: Bedrock age map.
         o Method: users can choose among 1D linear regression, 3D linear interpolation, 3D radial basis function. Alternatively, an externally produced age map can be imported.
2. Bedrock age uncertainty map interpolation
165   o Input: bedrock age-elevation dataset (point data) and bedrock age map (grid data).
         o Output: Bedrock age uncertainty map.
         o Method: The uncertainty of the age map is estimated through bootstrapping. An externally produced uncertainty map is required if the age map was imported.
3. Extract catchment bedrock age, coordinates, fertility and erosion data
170   o Input: catchment outline, bedrock age and age uncertainty maps, mineral fertility map, one or more erosion maps.
         o Output: a table of all the listed catchment properties necessary to predict detrital age distributions
         o Method: for each cell of the age map contained in the catchment outline, the local coordinates, age, fertility and erosional weight are extracted.
175 4. Predict detrital grain age distributions for each erosion scenario
         o Input: table of catchment data.
         o Output: a predicted detrital age population for each erosion scenario and related cumulative age distribution.
         o Method: A number of ages proportional to erosional weight and fertility is drawn from each cell, for each scenario. The ages from all catchment cells collectively represent a predicted population, from which
180   cumulative age distributions are constructed.
5. Evaluate the confidence of detecting erosion scenarios from a set of measured detrital ages
         o Input: one or more sets of observed grain-ages, the predicted detrital populations and distributions.

- o Output: A graph informing the confidence in discerning the tested scenarios from uniform erosion.

185
- o Method: predicted age distributions from erosion scenarios are compared to predicted uniform erosion distributions using the KS statistic as metric for dissimilarity. Confidence intervals are determined through Monte Carlo subsampling as a function of varying sample sizes.

6. Evaluate the fit of the erosion scenarios to the observed detrital ages
   - o Input: same as point 5).
   - o Output: two plots to visualize the fit of predictions to the observed age distribution(s).

190
   - o Method: in the first plot, dissimilarities calculated between predictions and observations are sorted and their distribution is plotted in the form of a violin plot. Secondly, a two-dimensional Multi-Dimensional Scaling (MDS) model is fitted to the dissimilarities among the predicted and observed distributions and plotted following Vermeesch (2013).

All of the above operations are embedded in an open-source Jupyter Notebook (jupyter.org), a software that allows integrating
195 text, Python code, and visualizations within the same document for maximum editability and transparency. All plots are produced with Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2020) Python libraries and are colored using the color blind-friendly and perceptually uniform ScientificColourMaps6 (Crameri et al., 2020). In the following paragraph we show how the program helps analyzing an already published detrital AHe age dataset (Stock et al., 2006).

5 Application of ESD_thermotrace to the Inyo Creek case study

200 We apply *ESD_thermotrace* to the bedrock and detrital apatite U-Th(-Sm)/He (AHe thereafter) age datasets of Stock et al. (2006). In this study, 9 bedrock AHe ages from the Inyo Creek and adjacent Lone Pine Creek catchments (eastern Sierra Nevada, California, USA) are used to constrain the age-elevation relationship of the source area. The authors compared these bedrock data to the AHe age distribution of river sand samples from both catchment outlets. For the sake of simplicity, here we only consider the fine sand sample (*k*=52) from the Inyo Creek catchment. Based on their analysis Stock et al. (2006) infer
205 no significant variability of sediment production (erosion) across all elevations within the catchment.

5.1 Data import and bedrock age interpolation

The first few steps of *ESD_thermotrace* allow importing (i) the digital elevation model of the study area, (ii) the polygon of the catchment of interest, (iii) a table of bedrock age and elevation, (iv) the (optional) erosion scenarios to be tested and (v) a (optional) fertility map. Next, the surface bedrock age is calculated with the preferred method (see above and README file
210 in Madella et al., 2021). The bedrock AHe ages of Stock et al. (2016), recalculated after Riebe et al. (2015), exhibit a ~60 My age increase from ca. 2 to 4.5 km of elevation range that is well-described by a linear regression ($R^2 = 0.94$). The input data are plotted in Figure 4a. In Figure 4b, every cell is assigned an AHe age as a linear function of elevation (Fig.5 inset), so as to map the bedrock cooling age on the topographic surface. The error of this interpolation is also mapped through bootstrapping
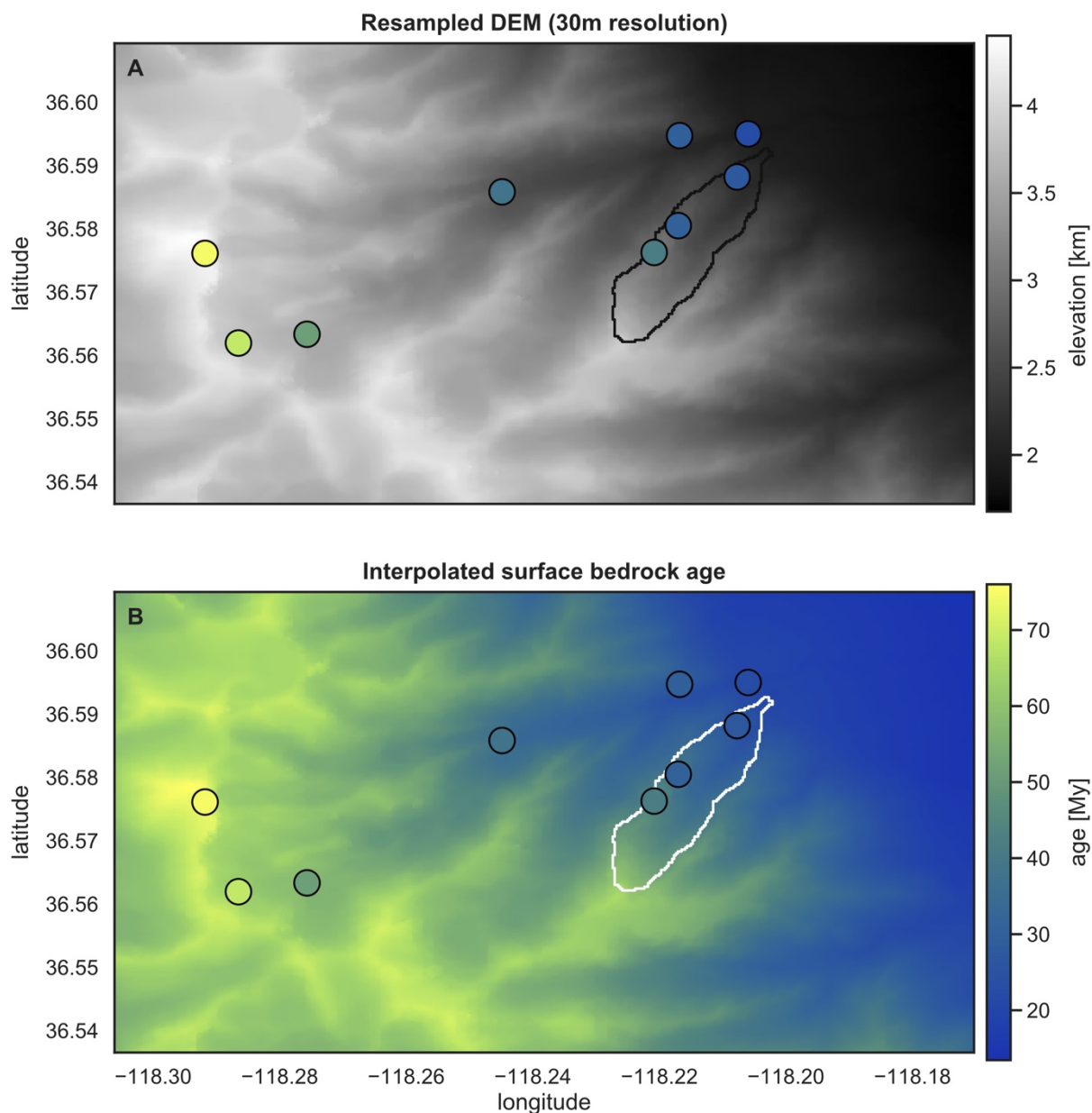
Figure 4 – Input bedrock data from Stock et al. (2006). (A) Raster images of the study area's digital elevation model, resampled to the user-specific cell size, and (B) bedrock surface AHe age interpolated based on the linear regression of age-elevation data. In both plots, point data inform bedrock sample locations and related AHe ages. Polygons shows the location of the Inyo Creek catchment.

as a function of elevation (Fig.5). This informs the elevation-dependent standard deviation of 1,000 regressions, in which the

215   bedrock ages are each drawn from a normal distribution that account for the input analytical uncertainty.
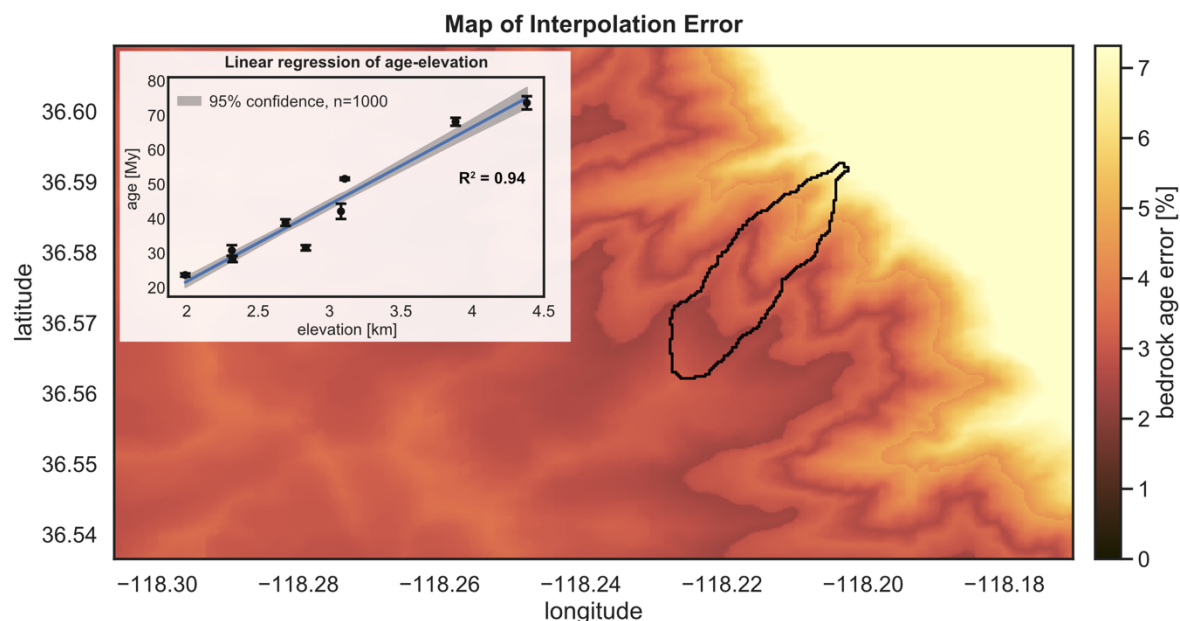
Figure 5 – The raster image displays the bedrock age interpolation error mapped to the topographic surface of the study area. The inset shows the linear regression employed as age-elevation function, as well as the associated 1σ uncertainty.

5.2 Extraction of catchment data

Next, *ESD_thermotrace* extracts the x,y,z coordinates, the bedrock cooling age and the related error for all the cells bound by the catchment outline. These data are written in a table, to which a column informing erosional weights is added for each desired erosion scenario and for mineral fertility. In addition to the user-defined erosion maps, by default the software considers

220 the uniform erosion scenario *'Euni'* (spatially constant erosional weight). Two further example scenarios can be toggled to test for exponential increase of erosion with elevation and an exponential decrease of erosion with elevation. We note that any other spatial variation in erosion can be defined by a user such that 'erosion maps' for the catchment following stream power, slope dependent, glacial slide velocity, or other approaches from geomorphic transport laws can be input. For this application, bulk geochemistry and point-counting analyses of Riebe et al. (2015) indicate that apatite fertility does not significantly vary

225 within the three lithologies found in the Inyo Creek catchment (Lone Pine granodiorite, Paradise granodiorite, Whitney granodiorites). For illustrative purpose, we test uniform erosion against two simple step-wise erosion functions of elevation: a scenario of 2-fold decrease in erosion below 3000 m ('*E<3000*') and the opposite scenario with 2-fold increase in erosion above 3000 m ('*E>3000*').

5.3 Prediction of age populations and age distributions

230 The erosional weights are used to forward model sediment production at each x,y position in the catchment. To do so, a number of grain-ages proportional to the erosional weight and to the local mineral fertility are randomly chosen for each cell. These

grain-ages are randomly drawn from a normal distribution constructed using the local interpolated bedrock age and the related error. The randomly picked grain-ages of all cells are stored in one suite of grain-ages, which represents the detrital grain-age population of a well-mixed fluvial sediment at the catchment outlet. Such a detrital population is forward modeled for each

235 erosion scenario, and related predicted CADs are constructed by sorting the age populations (Fig.6). For the sake of quick visual comparisons, *ESD_thermotrace* plots a kernel density estimation with arbitrary smoothing (Fig.6a). However, the
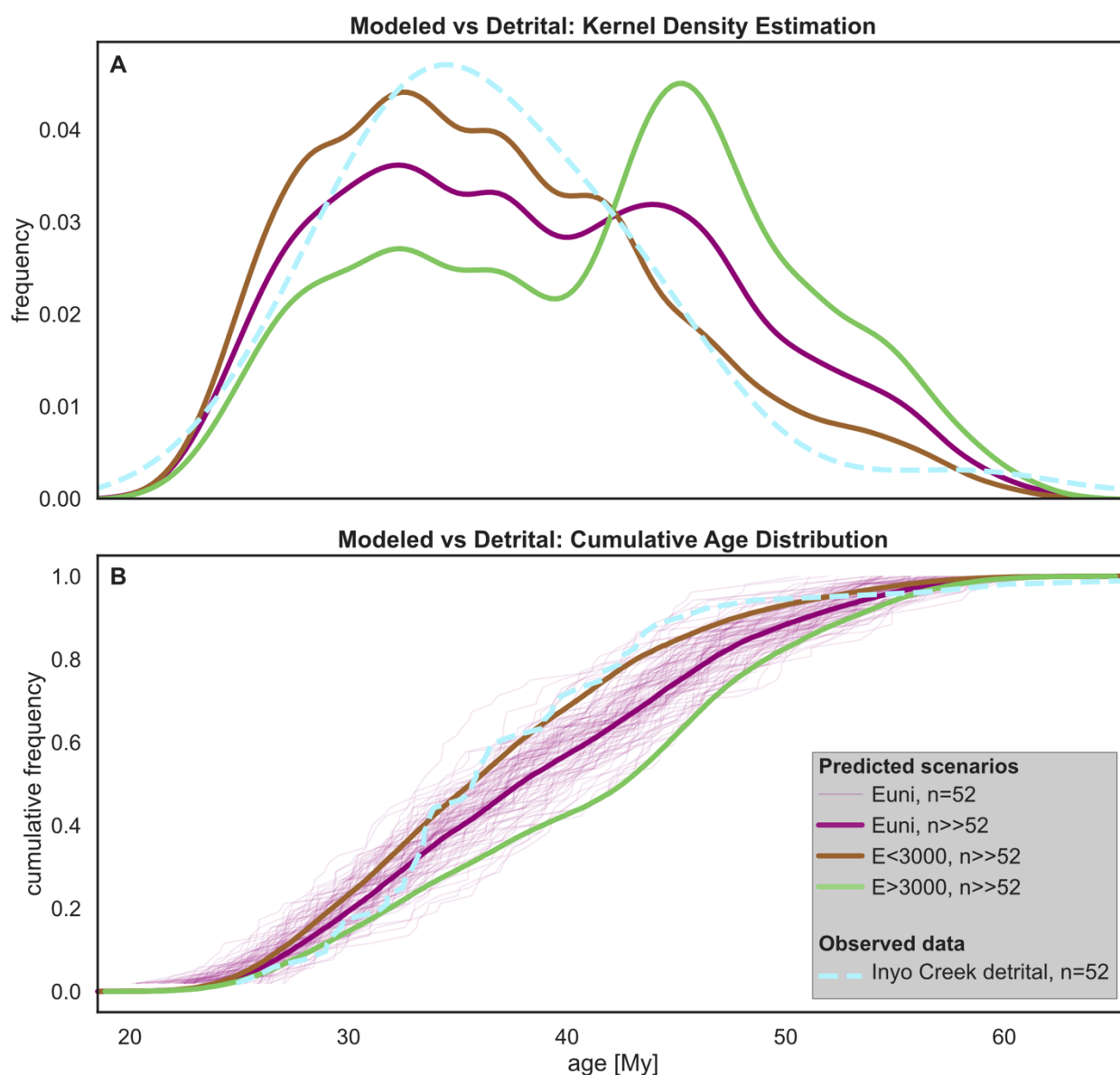

Figure 6 – Kernel Density Estimation curves of all predicted detrital distributions as well as of the observed detrital sample (A). Cumulative form of the same distribution, plus 100 random subsamples of the reference uniform erosion scenario (n=k=52), for a visual comparison. This plot can be customized to show the DKW-bounds as an alternative to random sampling. Euni: uniform erosion; E<3000: 2-fold erosion below 3000 m; E>3000: 2-fold erosion above 3000 m.

dissimilarity among distributions will be evaluated exclusively based on the cumulative form, in order to minimize bias (Vermeesch, 2007). In Figure 6b, all predicted and observed cumulative age distributions are plotted. For a first order impression of the uncertainty due to sample size, 100 $n=k$ random subsamples of the reference scenario are also plotted in the

240 background. Here, the user can specify for which reference scenario ("Euni" as default) the subsampled distributions should be plotted.

5.4 Confidence level as a function of sample size

After predicting the distributions, the software allows calculating the probability to discern each scenario from uniform erosion as a function of sample size. This is the main step to define how many grain-ages are necessary to detect a hypothetical erosion

245 pattern. The confidence in discerning scenarios is based on the iterative comparison between the reference scenario and all other predicted CADs. This is achieved through random subsampling of the predicted distributions for a range of possible sample sizes (30≤n≤140). For each n_i, 1,000 $n=n_i$ subsamples of the reference scenario are randomly drawn and the KS distance
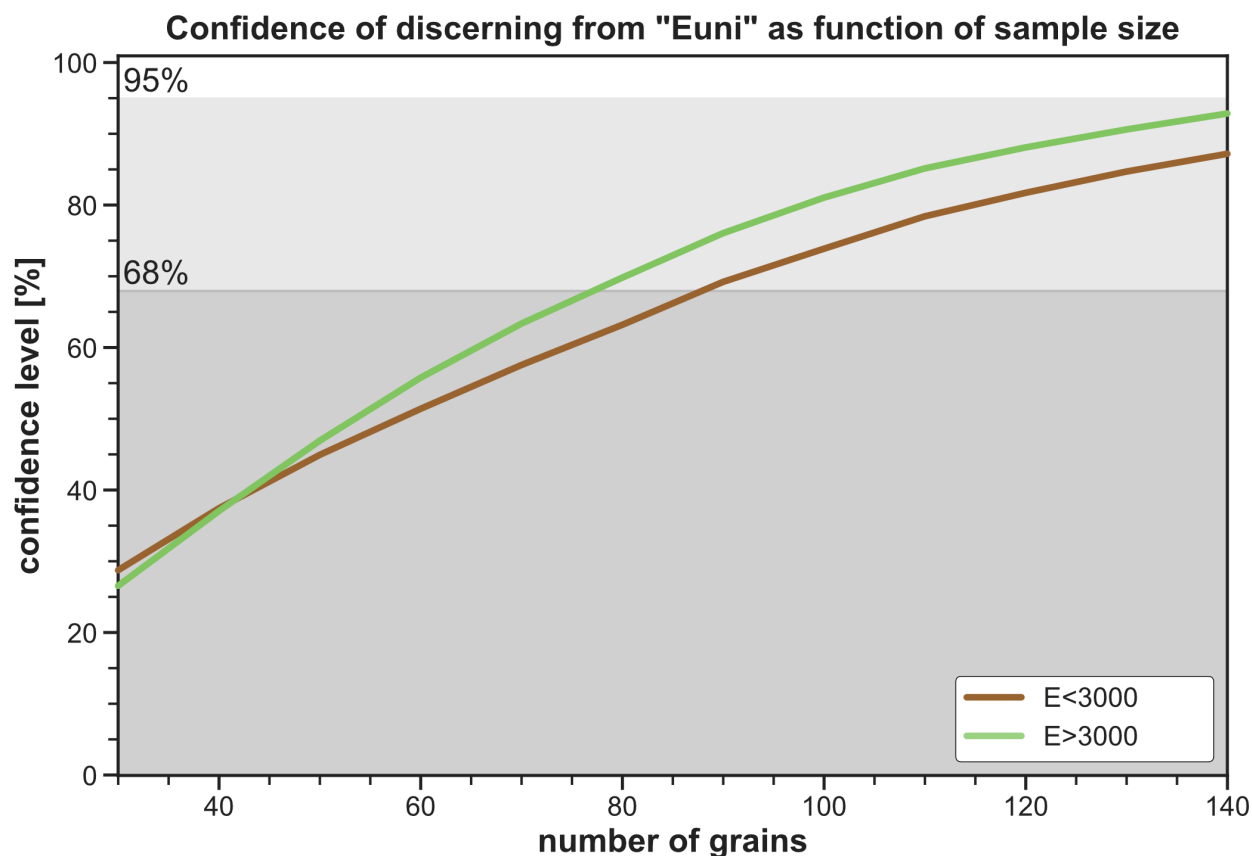

Figure 7 – Confidence level at which an erosion scenario can be discerned from the reference uniform erosion scenario, as function of the number of observation (i.e. number of detrital grain-ages). In this case study, among the two considered scenarios, 'E>3000' can be better distinguished from 'Euni'. Still, even 140 grain-ages would not suffice to reach the 95% confidence level. Euni: uniform erosion; E<3000: 2-fold erosion below 3000 m; E>3000: 2-fold erosion above 3000 m.

$d_{KS}(n_i)$ from the original $n>>n_i$ distribution is calculated for each iteration. The 95th percentile of all sorted $d_{KS}(n_i)$ is referred to as $d_{KS}(n_i)^*$, the inherent noise (i.e. dissimilarity) of the reference scenario for $n=n_i$. In the observations-cumulative frequency

250   space (Fig.6b), $d_{KS}(n_i)^*$ represents the vertical distance that any $n=n_i$ distribution needs to exceed in order to be discerned from a $n=n_i$ subsample of the reference scenario, with 95% confidence. Next, 1,000 $d_{KS}(n_i)$ are calculated between $n=n_i$ subsamples of each erosion scenario and the reference scenario. The probability of exceeding $d_{KS}(n_i)^*$ is plotted for each n_i (Fig.7). For the Inyo Creek case study (Stock et al., 2006) the tested erosion scenarios '$E<3000$' and '$E>3000$' yield predicted CADs that are not extremely dissimilar from uniform erosion. As a consequence, even 140 grain-ages would not suffice to detect these

255   scenarios at the 95% confidence level.

Figure 7 shows the use of *ESD_thermotrace* as a tool to explore the feasibility of a tracer thermochronology study. Based on their research question, users can import a number of possible erosion maps (or calculate them as functions of elevation) and test whether or not they can be detected through detrital tracer thermochronology. In other words, Figure 7 displays an

260   assessment of the number of grain-ages required to rigorously test a hypothesis. Such analysis of feasibility is not only beneficial in terms of scientific rigor, but it can also assist investigators in defining the budget for measurements at an early stage of proposal writing. Alternatively, in cases where the number of datable grains is limited by material properties, budget, or other logistic reasons, Figure 7 informs the confidence level of the results.

5.5 Evaluating the fit of each scenario to the detrital observations.

265   The final steps of *ESD_thermotrace* assist users in finding the scenario that best fits the observed distribution of detrital grain-ages. For each erosion scenario, d_{KS} and d_{Kui} (Fig.3) are calculated between 10,000 $n=k$ subsamples of the respective predicted CAD and the observed CAD. The distribution of these d_{KS} and d_{Kui} values is shown in the form of a split violin plot in Figure 8a and it is to be compared to the range of values shown by the violin in Figure 8b. The latter shows the distribution of d_{KS} and d_{Kui} calculated between random subsamples of the observed CAD that account for 1σ uncertainty, and the observed CAD itself

270   (constructed only with the mean ages). In other words, Fig. 8b visualizes a detrital distribution's noise due to analytical error and sample size. The plausibility of each scenario is plotted beneath each violin and is computed as the probability that the values plotted in Figure 8a fall within the on-sided lower 95th percentile of those shown in Figure 8b. More specifically, there is 62.9% fit to the observed CAD for the uniform erosion scenario, and a 87.6% and 3.3% fit for the factor of 2 increase in erosion below 3,000 m ('E<3000') and above 3,000 m ('E>3000'), respectively. Regardless of the best-fit erosion scenario,

275   the 52 available detrital grain-ages provide a (limited) confidence level of 45-50% for the interpretation. We refrain from making an improved interpretation of the erosion distribution in this catchment from that of Stock et al. (2006), because more grains would be needed, and as recent (and in progress) work has demonstrated, a larger range of different grain sizes should preferably be considered than that of Stock et al. (2006) to better understand the pattern of erosion (Lukens et al., 2019; see also discussion below; Riebe et al., 2015).
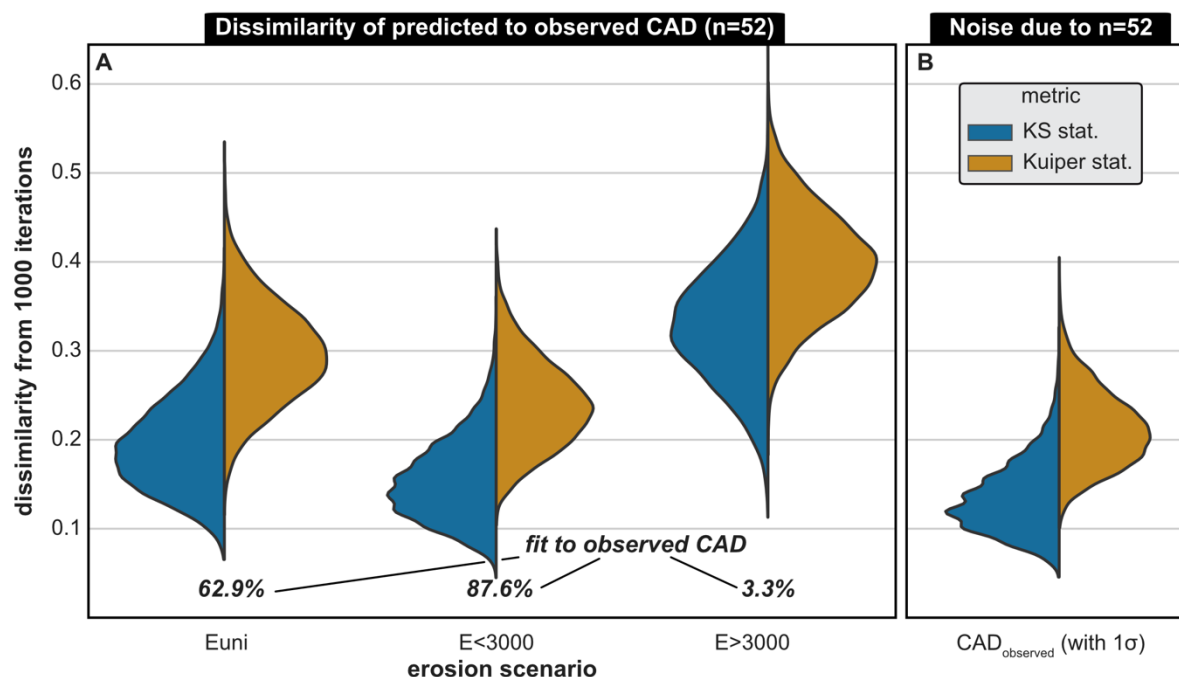
280

Figure 8 – Violin plot showing the KS (and Kuiper) statistics between the predicted CAD of each erosion scenario and observed CAD, calculated for 1000 n=k=52 subsamples (A). KS (and Kuiper) statistics between the observed CAD and 1000 random n=k=52 distributions, drawn from the same detrital age population but including uncertainty (B). Every violin is split in two halves with equal area, each representing the probability distribution function of 10000 KS (or Kuiper) statistics. The widest point of each semi-violin informs the median KS (or Kuiper) statistic; The closer to zero, the more similar is a scenario to the observed detrital age distribution. The most unlikely scenario is 'E>3000'. Both other scenarios are plausible, because their dissimilarity to CAD$_{observed}$ largely overlaps the range of dissimilarity due to analytical error of the detrital ages.

Lastly, as suggested by Vermeesch (2013), the *ESD_thermotrace* program also applies a two-components Multi-Dimensional Scaling model (MDS) to all predicted and observed distributions. This fits a coordinate system to the measured dissimilarities d_{KS} among all distributions. In a well-fit MDS model, distances calculated in the modelled 2D-space are a good approximation of the actual dissimilarities among the input elements (Fig.9a). To reach a satisfactory fit, modelled dissimilarities are plotted

285 against input dissimilarities (Fig.9b) and the sum of all distances to the 1:1 line is minimized (a procedure commonly referred to as *stress* minimization). The MDS plot renders an immediate visualization of the dissimilarity among distributions. Moreover, with the addition of the 68% and 95% confidence ellipses, the degree of overlap among distributions is also easily visually assessed (Fig.9a).

290 We have illustrated the functioning of the *ESD_thermotrace* by evaluating Stock et al.'s (2006) Inyo Creek data. Despite the good fit of scenario '*E<3000*', our analysis suggests that based on the used set of 52 grain-ages it is not possible to uniquely detect any of the tested scenarios at with confidence (Fig.7). This outcome stems from the considerable overlap between the range of possible predicted distributions '*Euni*' and the detrital observations (Fig.9). Nevertheless, using our approach, we add
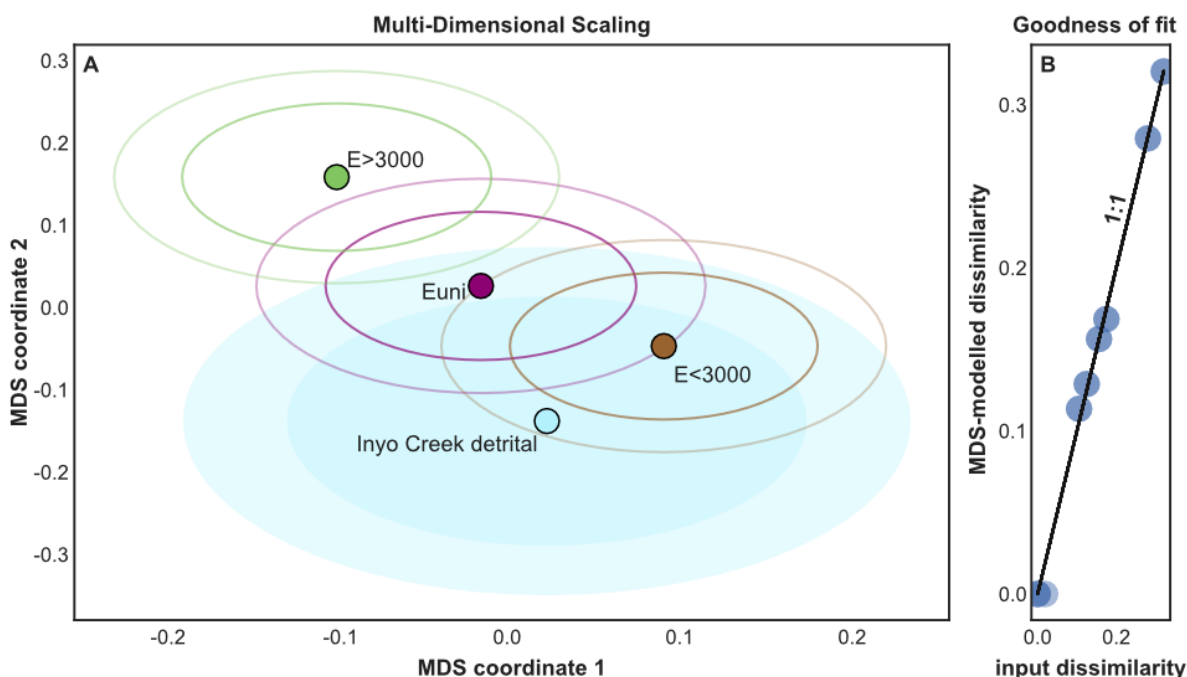
Figure 9 – Multi-Dimensional Scaling (MDS) plot. The 2-components MDS model fits a coordinate system to the measured dissimilarities (i.e. KS statistics) among all distributions. If the MDS model has a good fit, distances calculated in the new 2D-space (A) are a good approximation of the input dissimilarities among the elements. The goodness of fit is shown in (B), where modelled dissimilarities are plotted against input dissimilarities. For each distribution in (A) 68% and 95% confidence ellipses are also plotted.

a quantitative assessment of both the confidence due to sample size (45-50% confidence) and the scenarios fit. We note that

295   Riebe et al. (2015), in contrast, have shown that erosion in the Inyo Creek catchment is most likely to increase exponentially with elevation. These authors have fitted an erosional curve that additionally accounts for [10]Be-derived denudation rates, as well as 73 additional AHe ages from coarse sand-sized sediment. They conclude that grain size bias is a viable explanation of the observed discrepancy. While we will address the possible issues concerning grain size below, the best-fit erosional pattern of the Inyo Creek is beyond the scope of this paper and will not be further discussed.

300   6 How many grains do we need for tracer thermochronology?

The analysis with *ESD_thermotrace* shows that the appropriate sample size for a tracer thermochronology study cannot be determined a-priori without knowing (i) the case-specific scientific question (i.e. the spatial pattern of erosion to be tested), (ii) the source area hypsometry, (iii) the desired minimum confidence level and (iv) the surface bedrock ages. On this note, it is advisable to explore the feasibility of a study with already available bedrock ages before sampling. In absence of available

305   published ages, bedrock samples are best processed first in order to avoid wasting resources on potentially inconclusive analyses of detrital grains. To better illustrate the importance of the initial knowledge of the catchment's geology, we conducted a set of simulations with the same inputs as the Inyo Creek case study (Stock et al., 2006). In these simulations we vary the
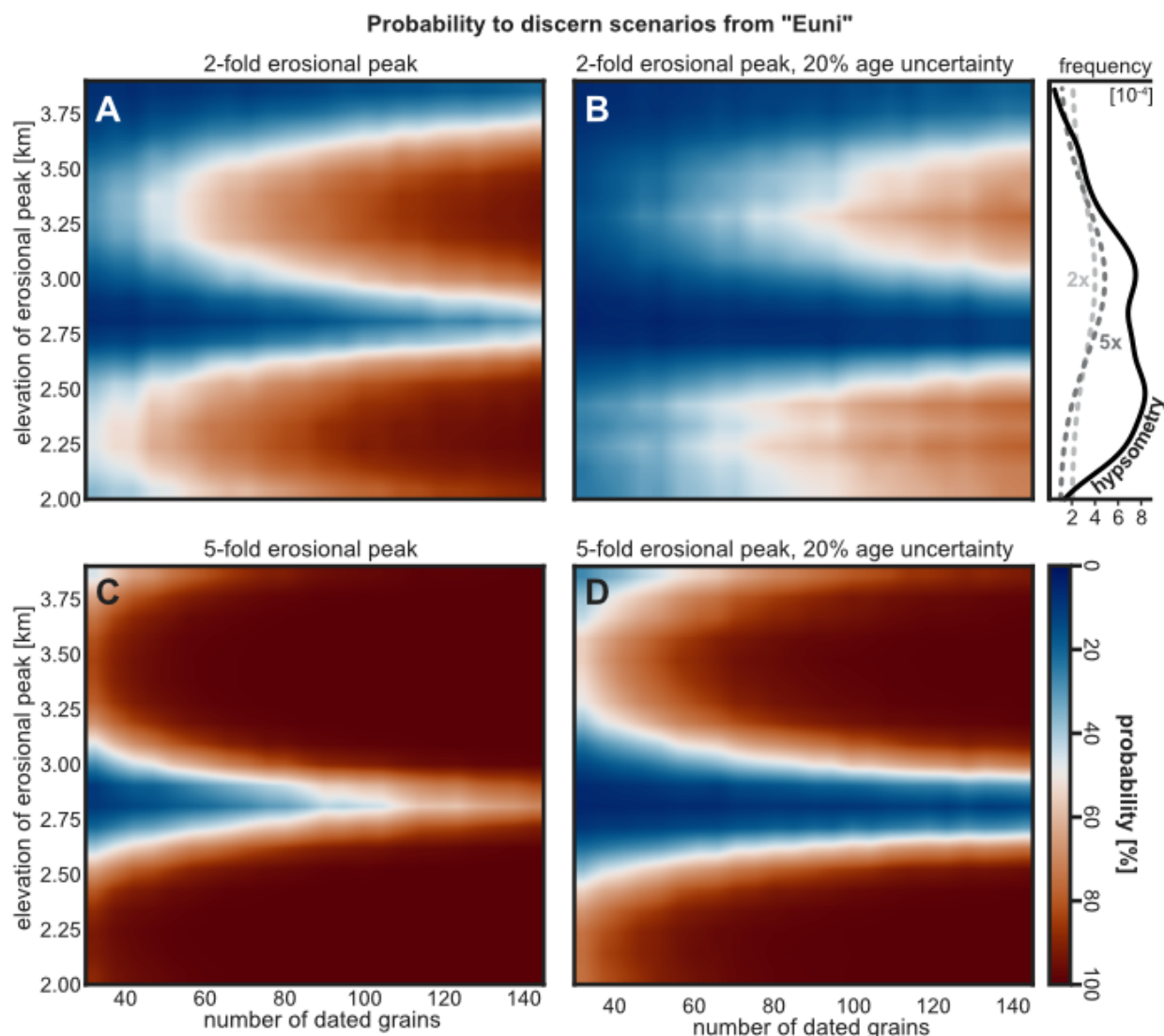
Figure 10 – Analysis of the confidence level in discerning between a detrital distribution generated with uniform erosion scenario 'Euni' and by a scenario with a gaussian peak of erosion with 1σ = 500 m. The color scale informs the probability as a function of the number of grains (x axis) and the elevation at which the gaussian peak is located (y axis). The catchment hypsometry (upper right corner) and age-elevation-relationship used here are taken from the Inyo Creek data (Stock et al., 2006). (A) Erosional function with peak 2 times greater than the catchment minimum, age uncertainty from Fig.5. (B) Erosional function with peak 2 times greater than the catchment minimum, 20% uncertainty on all bedrock ages. (C) Erosional function with peak 5 times greater than the catchment minimum, age uncertainty from Figure 5. (D) Erosional function with peak 5 times greater than the catchment minimum, 20% uncertainty on all bedrock ages. Example erosional functions for peaks at 3 km elevation are given in the upper right corner. The reader is referred to the text for more details.

location of maximal erosion and observe the impact on the confidence level in detecting the imposed pattern. A broad gaussian function of erosion (1σ = 500 m) is applied to the entire range of catchment elevations, shifting peak erosion in 100-m-steps

310   at each simulation. We test two sets of gaussian functions of elevation (2x and 5x), computed in such a way that peak erosion

equals 2 and 5 times the minimum found in the catchment, respectively. These scenarios are meant to simulate realistic geomorphologic situations, where erosion rates might experience a 2- or 5-fold increase within the same catchment. In addition, these scenarios are also tested setting all bedrock age uncertainties to 20%. Results from this parameter study and example erosional functions are illustrated in Figure 10.

315

Figure 10 shows how the confidence level in discerning these gaussian-peak-scenarios from spatially constant erosion ('*Euni*') is affected by both the elevation of the erosional peak, the amplitude of the peak, the sample size and the uncertainty of bedrock ages. An increase in sample size and/or an increase in amplitude of the erosional peak always correspond to higher success rates. In contrast, an increase in age uncertainty always reduces the success rate. The elevation of the erosional peak affects

320   the confidence level depending on the divergence from the catchment's hypsometric curve (Fig.10). Confidence minima (dark blue areas) are observed where the erosional peak is located at elevations greater than 3,500 m and at ca. 2,800 m (Fig. 10). The minima at 3,500 m occur because peak erosion involves an elevation range that is not frequent enough to produce a statistically relevant grain population. In contrast, the region of low confidence at around 2,800 m coincides with the broad frequency peak of catchment elevations (Fig.10). In this elevation range, the broad increase in erosional weights effectively

325   only accentuate the broader peak of the hypsometric curve. Remarkably, all detrital distributions resulting from peak erosion at ~2,800 m are statistically similar to those obtained from spatially uniform erosion, regardless of sample size. This implies that certain combinations of erosional pattern and distribution of bedrock ages are poorly suited to be investigated by means of tracer thermochronology.

330   Figure 11 displays an analogous parameter study to that illustrated in Figure 10, with the difference that all gaussian peaks of erosion are set to be narrower ($1\sigma = 100$ m). While Figure 11 maintains the same trends as Figure 10, it also shows that narrow, low-amplitude erosional peaks remain undetectable regardless of both peak elevation and sample size. However, high-amplitude, narrow erosional peaks (Fig.11C) are likely to be detected even in the critical elevation range of peak hypsometry individuated in Figure 10. The application of these simple erosional patterns to the parameter studies above demonstrate the

335   importance of analyzing the catchment hypsometry and of testing erosion scenarios even before collecting data, in order to make a science-informed choice on the appropriate sample size.

7 Other sources of uncertainty

The approach presented here for interpreting grain-age distributions enables the quantification of the confidence level in detecting erosion patterns with detrital tracer thermochronology. Confidence levels are computed as a function of the number

340   of samples, input erosion scenario and catchment properties. Nevertheless, even if a scenario is inferred to be highly probable, it may not be a unique solution or it may only inform the behavior of the one analyzed grain size fraction. Uncertainty in the interpretation can stem from factors such as: (1) complex bedrock age-elevation relationship and (2) spatial variability of sediment size resulting from transport distance (e.g. Lukens et al., 2019), geomorphic process (van Dongen et al., 2019; e.g.

Riebe et al., 2015), lithological differences (von Eynatten et al., 2012), or vegetation effects on weathering and erosion (Starke

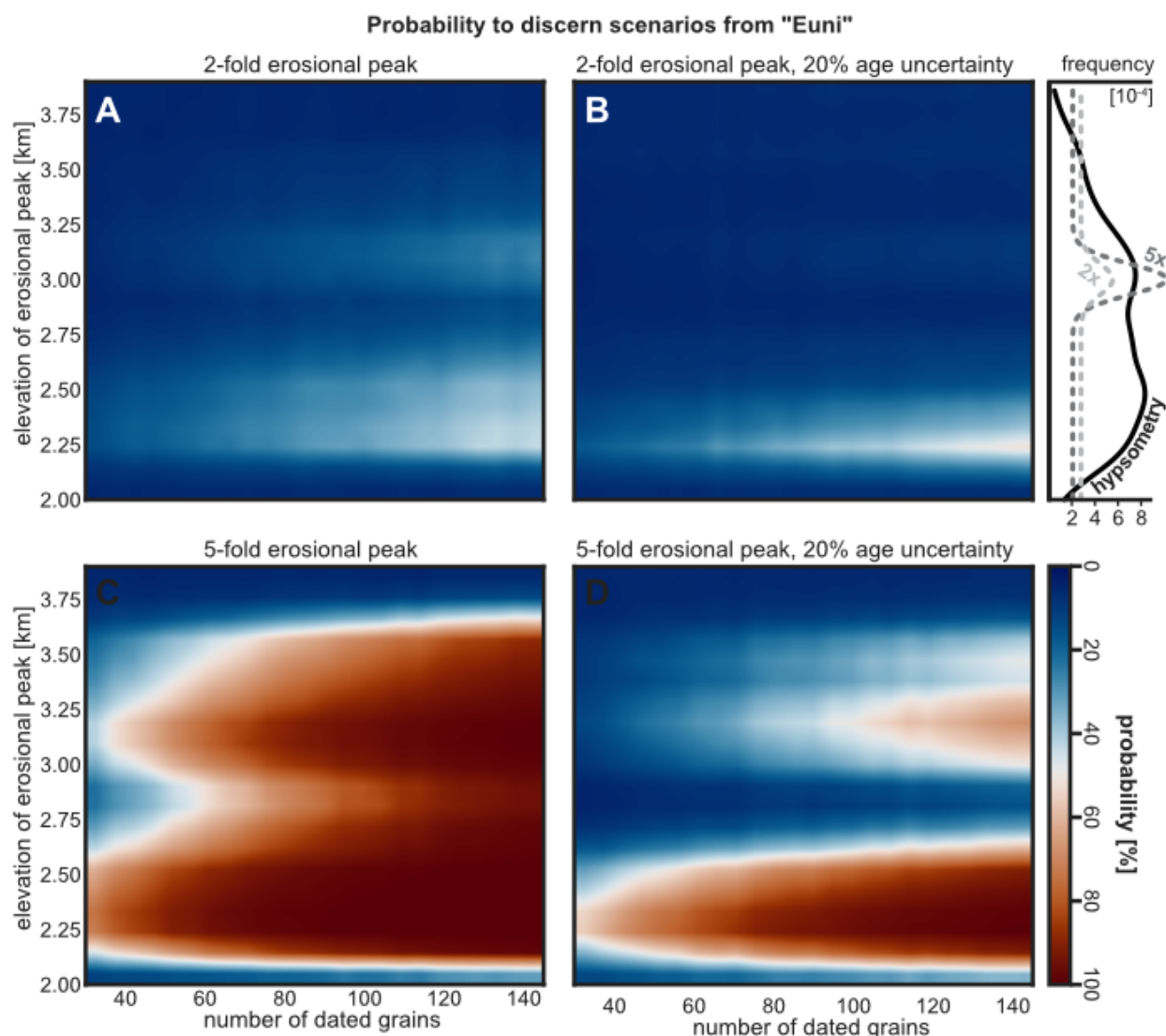345 et al., 2020). Both factors should be taken into account as discussed here below.

Figure 11 – Analysis of the confidence level in discerning between a detrital distribution generated with uniform erosion scenario 'Euni' and by a scenario with a gaussian peak of erosion with 1σ = 100 m. The color scale informs the probability as a function of the number of grains (x axis) and the elevation at which the gaussian peak is located (y axis). The catchment hypsometry (upper right corner) and age-elevation-relationship used here are taken from the Inyo Creek data (Stock et al., 2006). (A) Erosional function with peak 2 times greater than the catchment minimum, age uncertainty from Fig.5. (B) Erosional function with peak 2 times greater than the catchment minimum, 20% uncertainty on all bedrock ages. (C) Erosional function with peak 5 times greater than the catchment minimum, age uncertainty from Figure 5. (D) Erosional function with peak 5 times greater than the catchment minimum, 20% uncertainty on all bedrock ages. Example erosional functions for peaks at 3 km elevation are given in the upper right corner. The reader is referred to the text for more details.

18

In some cases, elevation alone cannot explain the entire variance of bedrock thermochronometric ages. For example, spatial variability of bedrock ages may reflect the proximity to tectonic structures or sub-catchment thermal events (e.g. magmatism, spatial changes in thermal gradients due to structures, etc). In such cases, an improved sampling of the spatial distribution of

350   bedrock ages is needed, and may require more complex interpolation functions (e.g. Glotzbach et al., 2013). To accommodate these complexities, *ESD_thermotrace* allows for 1D and 3D-linear interpolation as well as radial basis function interpolation. In all these cases, interpolated age uncertainty are calculated through bootstrapping. In addition, users can opt to import an independently interpolated surface bedrock map such as from a thermokinematic model (e.g. Whipp and Ehlers, 2019). For additional information on bedrock age mapping the reader is referred to the "README" file attached to the code (Madella et

355   al., 2021). Regardless of which method is used, users should be aware that depending on the quality of the bedrock data, different interpolation methods may yield different predicted distributions. Accordingly, resulting interpolated surfaces should be carefully evaluated, and a preference should be motivated by field observations and/or independent constraints. Lastly, the map of interpolated age uncertainty can inform the locations where additional bedrock sampling would help reduce the noise.

360   Other possible sources of bias concern the grain size of the analyzed samples. One issue is that downstream sediment abrasion may significantly modify detrital grain-age distributions, as can the weathering and erosion processes associated with a grain. For example, grains sourced the farthest from the sampling spot may be underrepresented in the analyzed grain size fraction (Lukens et al., 2019). A second issue is the initial grain size distribution found on the hillslopes. Depending on the lithology (von Eynatten et al., 2012) and on the locally dominant denudational process (van Dongen et al., 2019), different hillslopes of

365   the same catchment may produce significantly different sediment size distributions (e.g. Riebe et al., 2015). As a consequence, the mixed detrital sample may exhibit a bias in the relative abundance of the different age components. Both these issues can be mitigated through analysis of multiple grain size fractions (Lukens et al., 2019), multiple measures of hillslope sediment size distributions, composite analyses of trunk stream and tributary stream sediment samples, analyses of thermochronometers from different minerals.

370   **8 Conclusion**

This study reviewed previous approaches used to compare predicted and observed detrital grain-age distributions in the framework of tracer thermochronology. We have built upon these to develop a new tool (*ESD_thermotrace*) to investigate the upstream pattern of catchment erosion and the confidence level in uniquely inferring this as a function of sample size and study-site-specific variables. To demonstrate the utility of this approach, we presented an analysis of previously published data

375   from the Inyo Creek Catchment in California. The example highlighted the utility of measuring a large number of grains, and how multiple erosion scenarios are plausible for this catchment with the considered number of grains. The degree of statistical confidence permitted by this case study has also been quantified. We showed how the use of *ESD_thermotrace* can increase the statistical rigor of tracer thermochronology studies and how it can also assist investigators in budgeting analytical costs of

a future project. In cases where the number of datable grains is limited by any reason, the confidence level of the results can
380 be quantified.

Acknowledgements

Code availability

385 The source code of *ESD_thermotrace* is freely available for download (Madella et al., 2021). [Please note that a temporal preview access to the cited future DOI landing page is available: https://dataservices.gfz-potsdam.de/panmetaworks/review/c375e8eb059bb94ace622f2075b99baf6c513d3b560989a6b307bbe5d5a20217/]

References

Andersen, T.: Detrital zircons as tracers of sedimentary provenance: limiting conditions from statistics and numerical
390 simulation, Chem. Geol., 216(3–4), 249–270, https://doi.org/10.1016/j.chemgeo.2004.11.013, 2005.

Avdeev, B., Niemi, N. A. and Clark, M. K.: Doing more with less: Bayesian estimation of erosion models with detrital thermochronometric data, Earth Planet. Sci. Lett., 305(3–4), 385–395, https://doi.org/10.1016/j.epsl.2011.03.020, 2011.

Brewer, I. D., Burbank, D. W. and Hodges, K. V.: Modelling detrital cooling-age populations: Insights from two Himalayan catchments, Basin Res., 15(3), 305–320, 2003.

395 Clinger, A. E., Fox, M., Balco, G., Cuffey, K. and Shuster, D. L.: Detrital Thermochronometry Reveals That the Topography Along the Antarctic Peninsula is Not a Pleistocene Landscape, J. Geophys. Res. Earth Surf., 125(6), e2019JF005447, https://doi.org/10.1029/2019JF005447, 2020.

Crameri, F., Shephard, G. E. and Heron, P. J.: The misuse of colour in science communication, Nat. Commun., 11(1), 1–10, https://doi.org/10.1038/s41467-020-19160-7, 2020.

400 van Dongen, R., Scherler, D., Wittmann, H. and von Blanckenburg, F.: Cosmogenic 10Be in river sediment: where grain size matters and why, Earth Surf. Dyn., 7, 393–410, https://doi.org/10.5194/esurf-7-393-2019, 2019.

Ehlers, T. A., Szameitat, A., Enkelmann, E., Yanites, B. J. and Woodsworth, G. J.: Identifying spatial variations in glacial catchment erosion with detrital thermochronology, J. Geophys. Res. Earth Surf., 120(6), 1023–1039, https://doi.org/10.1002/2014JF003432, 2015.

405 Enkelmann, E. and Ehlers, T. A.: Evaluation of detrital thermochronology for quantification of glacial catchment denudation and sediment mixing, Chem. Geol., 411, 299–309, https://doi.org/10.1016/j.chemgeo.2015.07.018, 2015.

von Eynatten, H., Tolosana-Delgado, R. and Karius, V.: Sediment generation in modern glacial settings: Grain-size and source-rock control on sediment composition, Sediment. Geol., 280, 80–92, https://doi.org/10.1016/j.sedgeo.2012.03.008, 2012.

410 Glotzbach, C., van der Beek, P., Carcaillet, J. and Delunel, R.: Deciphering the driving forces of erosion rates on millennial to million-year timescales in glacially impacted landscapes: An example from the Western Alps, J. Geophys. Res. Earth Surf., 118(3), 1491–1515, https://doi.org/10.1002/jgrf.20107, 2013.

Glotzbach, C., Busschers, F. S. and Winsemann, J.: Detrital thermochronology of Rhine, Elbe and Meuse river sediment (Central Europe): implications for provenance, erosion and mineral fertility, Int. J. Earth Sci., 107(2), 459–479, https://doi.org/10.1007/s00531-017-1502-9, 2018.

415 Hunter, J. D.: Matplotlib: A 2D Graphics Environment, Comput. Sci. Eng., 9(3), 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Hurford, A. J. and Carter, A.: The role of fission track dating in discrimination of provenance, Geol. Soc. Lond. Spec. Publ., 57(1), 67–78, https://doi.org/10.1144/GSL.SP.1991.057.01.07, 1991.

Kuiper, N. H.: Tests concerning random points on a circle, Indag. Math. Proc., 63, 38–47, https://doi.org/10.1016/S1385-7258(60)50006-0, 1960.

Lang, K. A., Ehlers, T. A., Kamp, P. J. J. and Ring, U.: Sediment storage in the Southern Alps of New Zealand: New observations from tracer thermochronology, Earth Planet. Sci. Lett., 493, 140–149, https://doi.org/10/gdcf54, 2018.

Lukens, C. E., Riebe, C. S., Sklar, L. S. and Shuster, D. L.: Sediment size and abrasion biases in detrital thermochronology, Earth Planet. Sci. Lett., 115929, https://doi.org/10.1016/j.epsl.2019.115929, 2019.

425 Madella, A., Glotzbach, C. and Ehlers, Todd. A.: ESD_thermotrace, A new software to interpret tracer thermochronometry datasets and quantify related confidence levels, GFZ Data Services. https://doi.org/10.5880/fidgeo.2021.003, 2021.

Malusà, M. G., Carter, A., Limoncelli, M., Villa, I. M. and Garzanti, E.: Bias in detrital zircon geochronology and thermochronometry, Chem. Geol., 359, 90–107, https://doi.org/10.1016/j.chemgeo.2013.09.016, 2013.

Malusà, M. G., Resentini, A. and Garzanti, E.: Hydraulic sorting and mineral fertility bias in detrital geochronology, Gondwana 430 Res., 31, 1–19, https://doi.org/10.1016/j.gr.2015.09.002, 2016.

Massart, P.: The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality, Ann. Probab., 18(3), 1269–1283, https://doi.org/10.1214/aop/1176990746, 1990.

Massey, F. J.: The Kolmogorov-Smirnov Test for Goodness of Fit, J. Am. Stat. Assoc., 46(253), 68–78, https://doi.org/10.1080/01621459.1951.10500769, 1951.

435 McPhillips, D. and Brandon, M. T.: Using tracer thermochronology to measure modern relief change in the Sierra Nevada, California, Earth Planet. Sci. Lett., 296(3–4), 373–383, https://doi.org/10.1016/j.epsl.2010.05.022, 2010.

Nibourel, L., Herman, F., Cox, S. C., Beyssac, O. and Lavé, J.: Provenance analysis using Raman spectroscopy of carbonaceous material: A case study in the Southern Alps of New Zealand, J. Geophys. Res. Earth Surf., 120(10), 2056–2079, https://doi.org/10.1002/2015JF003541, 2015.

440 Riebe, C. S., Sklar, L. S., Lukens, C. E. and Shuster, D. L.: Climate and topography control the size and flux of sediment produced on steep mountain slopes, Proc. Natl. Acad. Sci., 201503567, https://doi.org/10.1073/pnas.1503567112, 2015.

Ruhl, K. W. and Hodges, K. V.: The use of detrital mineral cooling ages to evaluate steady state assumptions in active orogens: An example from the central Nepalese Himalaya: DETRITAL MINERAL COOLING AGES, Tectonics, 24(4), n/a-n/a, https://doi.org/10.1029/2004TC001712, 2005.

445 Spiegel, C., Siebel, W., Kuhlemann, J. and Frisch, W.: Toward a comprehensive provenance analysis: A multi-method approach and its implications for the evolution of the Central Alps, in Special Paper 378: Detrital thermochronology - Provenance analysis, exhumation, and landscape evolution of mountain belts, vol. 378, pp. 37–50, Geological Society of America, http://specialpapers.gsapubs.org/cgi/doi/10.1130/0-8137-2378-7.37, last access: 7 July 2016, 2004.

Starke, J., Ehlers, T. A. and Schaller, M.: Latitudinal effect of vegetation on erosion rates identified along western South 450 America, Science, 367(6484), 1358–1361, https://doi.org/10.1126/science.aaz0840, 2020.

Stock, G. M., Ehlers, T. A. and Farley, K. A.: Where does sediment come from? Quantifying catchment erosion with detrital apatite (U-Th)/He thermochronometry, Geology, 34(9), 725–728, 2006.

Vermeesch, P.: How many grains are needed for a provenance study?, Earth Planet. Sci. Lett., 224(3–4), 441–451, https://doi.org/10.1016/j.epsl.2004.05.037, 2004.

455 Vermeesch, P.: Quantitative geomorphology of the White Mountains (California) using detrital apatite fission track thermochronology, J. Geophys. Res., 112(F3), https://doi.org/10.1029/2006JF000671, 2007.

Vermeesch, P.: On the visualisation of detrital age distributions, Chem. Geol., 312–313, 190–194, https://doi.org/10.1016/j.chemgeo.2012.04.021, 2012.

Vermeesch, P.: Multi-sample comparison of detrital age distributions, Chem. Geol., 341, 140–146, 460 https://doi.org/10.1016/j.chemgeo.2013.01.010, 2013.

Waskom, M., Gelbart, M., Botvinnik, O., Joel Ostblom, J., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Warmenhoven, J., Cole, J. B., de Ruiter, J., Vanderplas, J., Hoyer, S., Pye, C., Miles, A., Swain, C., Meyer, K., Martin, M., Bachant, P., Quintero, E., Kunter, G., Villalba, S., Brian, B., Fitzgerald, C., Evans, C., Williams, M. L., O'Kane, D., Yarkoni, T. and Brunner, T.: mwaskom/seaborn: v0.11.1 (December 2020), Zenodo., 2020.

465 Whipp, D. M. and Ehlers, T. A.: Quantifying landslide frequency and sediment residence time in the Nepal Himalaya, Sci. Adv., 5(4), eaav3482, https://doi.org/10.1126/sciadv.aav3482, 2019.